# Habesha@DravidianLangTech: Abusive Comment Detection using Deep Learning Approach

**Mesay Gemeda Yigezu, Selam Abitte , Olga Kolesnikova, Grigori Sidorov, Alexander Gelbukh**

[∀] Instituto Politécnico Nacional (IPN), Centro de Investigación en Computación (CIC), Mexico

Correspondence: mgemedak2022@cic.ipn.mx

## Abstract

This research focuses on identifying abusive language in comments. The study utilizes deep learning models, including Long Short-Term Memory (LSTM) and Recurrent Neural Networks (RNNs), to analyze linguistic patterns. Specifically, the LSTM model, a type of RNN, is used to understand the context by capturing long-term dependencies and intricate patterns in the input sequences. The LSTM model achieves better accuracy and is enhanced through the addition of a dropout layer and early stopping. For detecting abusive language in Telugu and Tamil-English, an LSTM model is employed, while in Tamil abusive language detection, a word-level RNN is developed to identify abusive words. These models process text sequentially, considering overall content and capturing contextual dependencies.

## 1 Introduction

In recent years, online social networks (OSNs) have gained significant significance and have become a popular platform for obtaining news, information, and entertainment. However, despite the various advantages of utilizing OSNs, there is a mounting body of evidence indicating the presence of an escalating number of malicious individuals who exploit these networks to disseminate harmful content and cause damage to others. The negative consequences of these malicious activities are increasingly evident. The spread of poisonous content, such as hate speech, misinformation, and cyberbullying, can have severe psychological, emotional, and even physical effects on targeted individuals. Moreover, the virality and reach of OSNs amplify the potential harm caused by malevolent actors, as harmful content can quickly spread across networks, reaching a vast audience and causing widespread damage. In order to mitigate this activity the organizer provide this shared task.

Natural language processing (NLP) focuses on the practical manipulation of textual components, converting them into a format suitable for machines. Additionally, NLP plays a crucial role in Artificial Intelligence (AI) by providing vital insights to determine the positivity or negativity of information based on numerous comparisons. Hence we combat to fix the above-mentioned problems by applying the NLP concept.

## 2 Related Work

The objective of this study Chen et al. (2017) was to explore the use of core text mining techniques in automatically detecting abusive content in various social media platforms such as blogs, forums, media-sharing sites, Q and A platforms, and chat services. The research utilized datasets from popular platforms like Twitter, YouTube, MySpace, Kongregate, Formspring, and Slashdot. By employing supervised machine learning, the study compared different text representations and dimension reduction methods, including feature selection and feature enhancement. The results demonstrated the significant influence of these techniques on the accuracy of abusive content detection. Ultimately, the researcher concluded that employing a balanced dataset positively impacts the accuracy of detecting abusive content on social media platforms. They conducted to use minority class and majority class and obtain the best result on the minority class. In addition to that using feature reduction will improve efficiency whilst maintaining detection accuracies.

(Eshan and Hasan, 2017) investigates different machine learning algorithms to detect Bengali abusive text. After experiments, they analyzed that the SVM Linear kernel performs the best with trigram TfidfVectorizer features.

Awal et al. (2018) designed to detect abusive comments in social media by using Naïve Bayes. The researcher collected the corpus from YouTube. To calculate the occurrence of particular words in a particular comment used a bag of words (BOW)



*Third Workshop on Speech and Language Technologies for Dravidian Languages*,
pages 244–249, Varna, Bulgaria, Sep 7, 2023.
https://doi.org/10.26615/978-954-452-085-4_036

vector. In order to evaluate the performance of the model they applied 10-fold cross-validation.

The researcher (Akhter et al., 2021) undertook a comprehensive study focusing on the detection of abusive language in both Urdu and Roman Urdu comments. This investigation encompassed the utilization of a diverse set of machine learning and deep learning models. Specifically, the author employed five ML models, namely Naive Bayes, Support Vector Machine, Instance-Based Learning , Logistic Regression, and JRip. Additionally, four DL models, including CNN, LSTM, BLSTM, and Convolutional LSTM, were also harnessed in the analysis. The research methodology consisted of applying these models to two distinct datasets: a sizable collection comprising tens of thousands of Roman Urdu comments, and a comparatively smaller dataset containing over two thousand comments in Urdu. The primary objective was to assess the performance of these models in both linguistic variations. The outcomes of the experiments conducted revealed noteworthy insights. Notably, the CNN exhibited superior performance compared to the other models. Impressively, it achieved accuracy rates of 96.2% for Urdu comments and 91.4% for Roman Urdu comments. This marked the CNN as the most adept model in accurately identifying abusive language within these linguistic contexts.

This study (Emon et al., 2019) delves into the crucial task of identifying various forms of abusive content within the realm of online platforms. The research extensively explores the utilization of diverse machine learning and deep learning methodologies to address this challenge. The algorithms under scrutiny encompass a range of models, including the Linear Support Vector Classifier (LinearSVC), Logistic Regression (Logit), Multinomial Naïve Bayes (MNB), Random Forest (RF), Artificial Neural Network (ANN), and a RNN featuring aLSTM cell. This author introduces a pioneering dimension by devising novel stemming rules tailored specifically for the Bengali language. These rules substantially contribute to enhancing the efficacy and overall performance of the algorithms employed in the study. Notably, the deep learning-powered RNN model emerges as the frontrunner among the examined algorithms, boasting an impressive peak accuracy rate of 82.20%.

This scholarly investigation involves a comprehensive evaluation of the efficacy of Deep Learning (DL) models in contrast to Machine Learning (ML) models for the purpose of detecting instances of abusive language. The researcher, a notable contributor in this domain, undertook an empirical study wherein a comparative analysis was conducted between Logit and BLSTM models. The primary objective was to discern their effectiveness in identifying abusive language within the context of the Danish language. The study's findings showcase a distinct trend: the BLSTM model, a sophisticated variant of recurrent neural networks, emerged as the standout performer. It outshone the competing models in terms of accurately categorizing comments sourced from prominent social media platforms such as Reddit, Facebook, and Twitter.(Sigurbergsson and Derczynski, 2019) comprehensive experimentation and analysis revealed that the BLSTM model exhibited remarkable capabilities in dealing with the intricacies and nuances of abusive language present in user-generated content.

The studies discussed earlier have provided us with a valuable insight, demonstrating that the utilization of deep learning methodologies offers a straightforward means of detecting the intended content. What's even more advantageous about employing deep learning is that it eliminates the necessity for employing supplementary feature extraction techniques. This implies that the inherent capability of deep learning models allows them to discern patterns and features directly from the data, obviating the need for manual feature engineering. The integration of deep learning into content detection thus emerges as a pivotal advancement in the field, heralding a new era of intelligent and efficient detection mechanisms. This transformative approach holds promise for a wide array of applications where accuracy and automation are paramount.

## 3 Task description

The objective of the task is to determine whether a comment includes any form of abusive content. The data sets consist of YouTube comments written in the Tamil and Telugu-English languages. The comments or posts in the corpus can consist of multiple sentences, but the average sentence length of the corpus is one. The annotations in the corpus are made at the comment or post level, rather than at the sentence level. This means that the task involves analyzing comments or posts as a whole to determine if they contain abusive content, rather

than focusing on individual sentences within the comment. The annotations or labels indicating whether a comment is abusive or not are assigned based on the overall content of the comment or post (Priyadharshini et al., 2023).

## 4 Methodology

### 4.1 Data pre-processing

After receiving the data from shared task organizer (Priyadharshini et al.), it was split into three distinct parts: the training set, the development set, and the testing set. However, prior to utilizing this data for training purposes, it is crucial to perform pre-processing on it. The data is currently in an unsuitable format for training a model effectively. Therefore, it requires transformation into a readable and structured format that aligns with the requirements of the training process.

One of the primary tasks during pre-processing is the removal of various unwanted elements present in the data. These elements include links, HTML tags, numbers, and symbols that may hinder the training process or introduce noise into the dataset. By eliminating these unwanted components, the data becomes cleaner and more focused, enabling the model to better discern patterns and relationships within the text.

Once these unwanted elements have been removed, the data will be better suited for training the model. Pre-processing allows the model to focus on the relevant linguistic features and patterns within the text, improving its ability to generalize and make accurate predictions or classifications.

By performing the necessary pre-processing steps, such as transforming the data into a readable format and removing unwanted elements, the dataset will be optimized for training the model, facilitating more accurate and meaningful results in subsequent analysis or applications. The data size provided by the organizers of the shared task is visually represented in Figure 1.

### 4.2 Algorithms

Within this section, we delve into the algorithms employed within this research paper. The realm of linguistic modeling predominantly relies on the utilization of deep learning models (Yigezu et al., 2021; Arif et al., 2022). These models, such as Convolutional Neural Networks (CNN) and RNN, are commonly employed due to their ability to identify intricate patterns within textual data. More

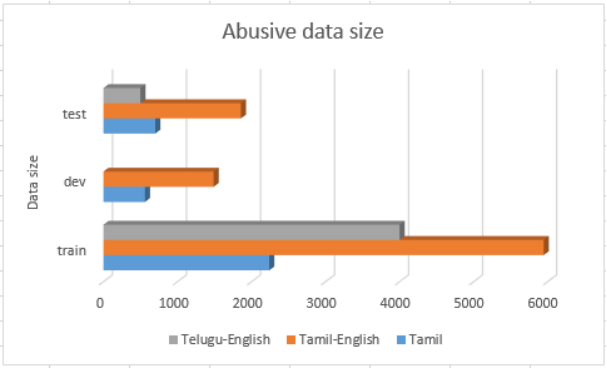


Figure 1: data set size for abusive comment detection

specifically, the LSTM model, which exhibits a tree-like structure, is employed as a recurrent neural network to effectively analyze sequential data of varying lengths (Yigezu et al., 2022).

To detect Telugu and Tamil-English abusive language, an LSTM model was employed. In the implementation, a dropout layer was added after the RNN layer. This dropout layer aids in mitigating the risk of overfitting, which can occur when the model excessively learns from the training data, leading to reduced generalization performance. Additionally, the model was configured to utilize early stopping based on validation loss. This mechanism halts the training process if the validation loss fails to exhibit improvement for a specified number of epochs. This approach helps prevent unnecessary computational effort and ensures that the model is not trained beyond a point where it ceases to benefit from further iterations. The LSTM model employed in our experiment exhibits better accuracy and significantly improves the contextual understanding of the data. By leveraging the capabilities of the LSTM, our model is able to effectively capture long-term dependencies and intricate patterns within the input sequences, enabling a deeper comprehension of the data.

Table 1 shows a comprehensive overview of the parameters utilized in our LSTM model. These parameters, carefully selected and fine-tuned, play a crucial role in shaping the model's architecture and optimizing its performance. By configuring the LSTM model with the appropriate parameters, we were able to enhance its ability to process sequential data and achieve notable accuracy in identifying abusive language.

In the context of Tamil abusive language detection, we developed and trained a rudimentary RNN at the word level to effectively identify abusive

| Parameters | Values |
|---|---|
| embed_units | 100 |
| hidden_units | 128 |
| dropout | 0.5 |
| optimizer | adam |
| batch_size | 64 |
| loss | categorical_crossentropy |
| epoch | 25 |
| activation | softmax |
| restore best weights | True |

Table 1: parameters used in LSTM

words. Word-level RNNs process text by considering words as a sequential input, generating predictions and hidden states at each step, and forwarding the most recent hidden state to the subsequent step. This iterative process allows the model to capture contextual dependencies and patterns within the sequence of words. RNNs have been widely employed as fundamental components in contemporary neural networks designed for language identification tasks.

To facilitate the mapping of tokens (i.e., words) to numerical representations, we utilized the Dictionary class. This class serves as a tool to assign unique and consecutive integer indexes to each token in the vocabulary. By mapping tokens to indexes, the model can efficiently handle text data and perform computations based on these numerical representations. This enables the RNN model to process and analyze the textual information effectively, aiding in the identification of abusive language.

To ensure fair and unbiased predictions, we employed a balanced dataset during our experimentation. By using a balanced dataset, we aimed to mitigate any potential bias that may arise from an imbalanced distribution of abusive and non-abusive instances. A balanced dataset consists of an equal number of instances from each class, which helps to prevent the model from favoring one class over the other during training and prediction.

By utilizing a balanced dataset, we strived to create a more equitable and reliable predictive model. This approach allows the model to learn from an unbiased representation of both abusive and non-abusive language, enhancing its ability to generalize and make accurate predictions on unseen data. Ultimately, the use of a balanced dataset contributes to the fairness and integrity of the abusive language detection system we developed. Table 2 depicts the parameters which we used in this experiment.

| Parameters | Values |
|---|---|
| embedding_size | 100 |
| hidden_size | 128 |
| optimizer | adam |
| batch_size | 32 |
| crossentropy loss | reduction = sum |
| num_iteration | 30 |

Table 2: parameters used in RNN

We used PyTorch, a popular deep learning framework. It creates an instance of the CrossEntropyLoss class from the torch.nn module and sets the reduction parameter to 'sum'. Torch.nn.CrossEntropyLoss is a loss function commonly used for multi-class classification problems. It combines the softmax function and the negative log-likelihood loss. It expects the input logits (unnormalized scores) and the target labels. The reduction parameter determines how the loss is aggregated over the batch. In this case, 'sum' means that the loss values for each element in the batch will be summed together to produce a single scalar loss value.

## 5 Result and Discussion

According to the findings presented in Table 3, we employed an RNN model to detect Tamil abusive languages. The evaluation results indicated that the RNN model achieved a precision of 0.26, a recall of 0.23, and an F1 score of 0.22. These metrics provide insights into the model's performance, with precision representing the accuracy of positive predictions, recall indicating the model's ability to correctly identify positive instances, and the F1 score representing the harmonic mean of precision and recall.

In the second experiment, our primary objective was to detect Tamil-English and Telugu-English languages using an LSTM model. The results demonstrated that the LSTM model performed better in the Telugu-English detection task, achieving a precision of 0.65, a recall of 0.65, and an F1 score of 0.65. On the other hand, for the Tamil-English detection, the LSTM model achieved a precision of 0.27, a recall of 0.25, and an F1 score of 0.26.

The performance metrics offer valuable insights into how effectively the models can identify in-

stances of abusive language in Tamil-English and Telugu-English texts. The higher precision, recall, and F1-score in the Telugu-English detection task indicate that the LSTM model exhibited superior performance in accurately identifying abusive language in that language pair. However, it is important to note that the model's performance was relatively lower in the Telugu-English detection task, suggesting the need for further refinements and improvements in the model's ability to identify abusive language in that specific language pair.

| Task | Macro-score | | | |
|---|---|---|---|---|
| | P | R | F1 | Acc |
| Tamil | 0.26 | 0.23 | 0.22 | 0.46 |
| Telugu-English | 0.65 | 0.65 | 0.65 | 0.65 |
| Tamil-English | 0.27 | 0.25 | 0.26 | 0.51 |

Table 3: Experimental results

## 6 Conclusion

The research paper utilizes deep learning models, specifically LSTM and RNNs, to analyze language patterns. The LSTM model is employed to enhance contextual understanding by capturing long-term dependencies and intricate patterns in the input sequences. When detecting abusive language in Telugu and Tamil-English, an LSTM model is utilized, resulting in improved accuracy and a deeper understanding of the context. In the context of Tamil abusive language detection, a word-level RNN is created and trained to identify abusive words. This type of RNN processes text sequentially, capturing contextual dependencies and patterns within the sequence of words.

To attain enhanced and superior performance levels, our focus will center on the utilization of a transformer-based approach (Aurpa et al., 2022; Gupta et al., 2022). Our strategy involves intricately configuring the parameters of this approach, meticulously fine-tuning them to yield outcomes that hold great promise. This deliberate effort aims to extract the utmost potential from the transformer model, thereby optimizing its performance for optimal results. Through this approach, we intend to unlock new dimensions of efficiency and effectiveness, pushing the boundaries of achievement in our pursuit of excellence.

## Acknowledgments

The work was done with partial support from the Mexican Government through the grant A1S-47854 of CONACYT, Mexico, grants 20220852, 20220859, and 20221627 of the Secretaría de Investigación y Posgrado of the Instituto Politécnico Nacional, Mexico. The authors thank the CONACYT for the computing resources brought to them through the Plataforma de Aprendizaje Profundo para Tecnologías del Lenguaje of the Laboratorio de Supercómputo of the INAOE, Mexico and acknowledge the support of Microsoft through the Microsoft Latin America Ph.D. Award.

## References

Muhammad Pervez Akhter, Zheng Jiangbin, Irfan Raza Naqvi, Mohammed AbdelMajeed, and Tehseen Zia. 2021. Abusive language detection from social media comments using conventional machine learning and deep learning approaches. *Multimedia Systems*, pages 1–16.

Muhammad Arif, Atnafu Lambebo Tonja, Iqra Ameer, Olga Kolesnikova, Alexander Gelbukh, Grigori Sidorov, and Abdul Gafar Manuel Meque. 2022. Cic at checkthat! 2022: multi-class and cross-lingual fake news detection. *Working Notes of CLEF*.

Tanjim Taharat Aurpa, Rifat Sadik, and Md Shoaib Ahmed. 2022. Abusive bangla comments detection on facebook using transformer-based deep learning models. *Social Network Analysis and Mining*, 12(1):24.

Md Abdul Awal, Md Shamimur Rahman, and Jakaria Rabbi. 2018. Detecting abusive comments in discussion threads using naïve bayes. In *2018 International Conference on Innovations in Science, Engineering and Technology (ICISET)*, pages 163–167. IEEE.

Hao Chen, Susan Mckeever, and Sarah Jane Delany. 2017. Harnessing the power of text mining for the detection of abusive content in social media. In *Advances in Computational Intelligence Systems: Contributions Presented at the 16th UK Workshop on Computational Intelligence, September 7–9, 2016, Lancaster, UK*, pages 187–205. Springer.

Estiak Ahmed Emon, Shihab Rahman, Joti Banarjee, Amit Kumar Das, and Tanni Mittra. 2019. A deep learning approach to detect abusive bengali text. In *2019 7th International Conference on Smart Computing & Communications (ICSCC)*, pages 1–5. IEEE.

Shahnoor C Eshan and Mohammad S Hasan. 2017. An application of machine learning to detect abusive bengali text. In *2017 20th International conference of computer and information technology (ICCIT)*, pages 1–6. IEEE.

Vikram Gupta, Sumegh Roychowdhury, Mithun Das, Somnath Banerjee, Punyajoy Saha, Binny Mathew, Animesh Mukherjee, et al. 2022. Multilingual abusive comment detection at scale for indic languages. *Advances in Neural Information Processing Systems*, 35:26176–26191.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Chinnaudayar Navaneethakrishnan, Malliga Subramanian, Kogilavani Shanmugavadivel, Premjith B, Abirami Murugappan, Sai Prashanth Karnati, Rishith, Chandu Janakiram, and booktitle = Kumaresan, Prasanna Kumar". Findings of the shared task on Abusive Comment Detection in Tamil and Telugu.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Malliga Subramanian, Subalalitha Chinnaudayar Navaneethakrishnan, Kogilavani Shanmugavadivel, Premjith B, Abirami Murugappan, Prasanna Kumar Kumaresan, Karnati Sai Prashanth, Mangamuru Sai Rishith Reddy, and Janakiram Chandu. 2023. Overview of shared-task on abusive comment detection in tamil and telugu. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, Varna, Bulgaria. Recent Advances in Natural Language Processing.

Gudbjartur Ingi Sigurbergsson and Leon Derczynski. 2019. Offensive language and hate speech detection for danish. *arXiv preprint arXiv:1908.04531*.

Mesay Gemeda Yigezu, Atnafu Lambebo Tonja, Olga Kolesnikova, Moein Shahiki Tash, Grigori Sidorov, and Alexander Gelbukh. 2022. Word level language identification in code-mixed kannada-english texts using deep learning approach. In *Proceedings of the 19th International Conference on Natural Language Processing (ICON): Shared Task on Word Level Language Identification in Code-mixed Kannada-English Texts*, pages 29–33.

Mesay Gemeda Yigezu, Michael Melese Woldeyohannis, and Atnafu Lambebo Tonja. 2021. Multilingual neural machine translation for low resourced languages: Ometo-english. In *2021 International Conference on Information and Communication Technology for Development for Africa (ICT4DA)*, pages 89–94. IEEE.